DOCUMENT RESUME

ED 462 116 JC 020 184

AUTHOR	Martinez, Daniel
TITLE	Predicting Student Outcomes Using Discriminant Function Analysis.
PUB DATE	2001-05-00
NOTE	22p.; Paper presented at the Annual Meeting of the Research and Planning Group (39th, Lake Arrowhead, CA, May 2-4, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; *Community Colleges; *Discriminant Analysis; *Educational Research; Evaluation Methods; Multiple Regression Analysis; *Outcomes of Education; Predictor Variables; Two Year Colleges

ABSTRACT

This document discusses how community colleges often under-utilize collected student data. For example, community colleges use student pre-college assessment data to place students in specific courses. The paper suggests that student data can also be used to predict academic success (a grade of A, B, or C) in community college courses. Some of the limitations of using multiple regression for decisions on student placement are also discussed. The use of discriminant function analysis is presented as an alternative method to help make more informed decisions on placement as well as predict possible student success in specific courses. A detailed explanation of how discriminant function analysis can be designed and used by community college researchers is provided. The report also includes a literature review of relevant research on the topics of research designs and student academic success. (Contains 23 references and 6 sample tables that display how discriminant function analysis results may be presented and interpreted.) (MKF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 116

Predicting student outcomes using discriminant function analysis

Daniel Martinez, Ph.D.
Riverside Community College District

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
D. Martinez
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Running head: Predicting Student Outcomes

JC 020184

BEST COPY AVAILABLE

Abstract

Most California community colleges collect copious amounts of data on entering students, most often through the assessment process. However, many times, the data are underutilized: only a few of the data elements captured are used for assessment purposes and the data are note used outside of placement. I have made several attempts to utilize the data, including an attempt to identify variables that would predict success in specific courses using multiple regression. Though this technique can be used to develop models to predict future behavior, it proved to be unfit for helping place students in courses because it can only be used to develop models based on success in a course, not placement into the course. Discriminant function analysis can provide the necessary classification into courses, though the development of a predictive model can prove intimidating. This research explores the limitations of using multiple regression for placement, the use of discriminant function as an alternative, and one method for using discriminant function to provide a model of future behavior.

Predicting student outcomes using discriminant function analysis

Many research questions in education seek to predict student outcomes based upon a set of independent variables. These variables may include high school information, background information, or scores on a test. Predicting student outcomes is really a process of trying to determine what group an individual student belongs. Should the student be placed into English 1A or a developmental English course? Will the student be more likely to dropout or be put on probation due to poor academic performance during their first semester? Reliable answers to these questions, and others like them, could help colleges tailor services and interventions to target populations and thereby utilize their limited resources more efficiently.

The method by which these predictions are made is usually by some statistical technique such as multiple regression. Multiple regression is used in a wide range of applications in social science research (Schroeder, Sjoquist, & Stephan, 1986) and was the initial method of analysis for the research that inspired this paper. However, multiple regression is best used when the outcome, or more generally, the DV, is either dichotomous or interval data (although "with appropriate coding, any comparison can be represented" [Cohen & Cohen, 1983, p. 512]). In the following scenario, I will described my use of multiple regression, the problem I encountered while created a model, and my ultimate decision to use Discriminant Function Analysis, a decision that ultimately proved the most helpful to the problem at hand.

Literature review

College admissions processes often depend on the ability to predict student success. However, the use of a test to help determine admission has traditionally been problematic and continues to be so. Recently, the chancellor of the University of California called for the end of using testing for admissions to college (Selingo & Brainard, 2001). This was not a new call: a

plethora of research has shown that standardized tests do not predict success equally well for all groups (Cleary, Humphreys, Kendirick, & Wesman, 1975; Melnick, 1975; Nettles, Thoeny, & Gosman, 1986; Tracey & Sedlacek, 1985) and that standardized tests do not measure what they claim to measure (Riehl, 1994; Sturm & Guinier, 2001). In a recent issue of Boston Review (2001), Susan Sturm and Lani Guiner attack the use of standardized tests in defense of affirmative action, stating:

> [W]e dispute the notion that merit is identical to performance on standardized tests. Such tests do not fulfill their stated function. They do not reliably identify those applicants who will succeed in college or later in life, nor do they consistently predict those who are most likely to perform well in the jobs they will occupy (p. 4).

As an alternative to standardized tests, Strum and Guiner suggest the use of multiple measures as a better way of deciding entry into law school.

Often, colleges may rely on two tests as a means of using multiple criteria, but if the two tests are highly correlated with each other, there is needless duplication in measuring the same aspect of a construct (Anastasi, 1982). Because the use of standardized tests has been shown to be problematic, multiple selection methods are being used to predict student success (Ebmeir & Schmulbach, 1989). The use of using multiple measures is called triangulation, the goal of which is to "strengthen the validity of the overall findings through congruence and/or complementarity of the results of each method" (Greene & McClintock, 1985, p. 524). This method is used extensively in education for admissions (Markert & Monke, 1990; McNabb, 1990) and involves using a variety of techniques simultaneously to measure a student's knowledge, skills, and values (Ewell, 1987).

Colleges can benefit from combining cognitive and noncognitive variables in predicting student academic success (Young & Sowa, 1992). Because the essence of triangulation is to

measure the same construct in independent ways (Greene & McClintock, 1985), the more non-related information gathered, the better the prediction. Triangulation can also minimize or decrease the bias inherent in any particular method by counterbalancing another method and the biases inherent in the other method (Mathison, 1988). For instance, most researchers rely heavily on survey research; however, the assumptions of survey research (e.g., the survey asked all the pertinent questions in a format the respondent can understand) are usually never questioned as a study is designed (Stage & Russell, 1992) which may lead to incomplete or inaccurate conclusions.

In the California Community Colleges, the required assessment process dictates the use of multiple measures in placing students into courses. Though the use of a test as one of the multiple measures is highly regulated, the use of multiple measures is not – unless using another test. Because of this, most multiple measures are chosen based on anecdotal or gut reactions and rarely on statistical evidence. It is the lack of research-based decisions for using multiple measures that inspired this research.

Collecting data and building a model

Many colleges collect more data than they use for analysis on a regular basis. Some examples of data captured from students as a part of assessment include:

- Age
- Ethnicity
- Sex
- English as the primary language
- Disability
- Admission status
- Veteran status
- High school education
- Highest degree earned
- Years out of high school
- Years of high school English
- Grade in last English course
- High school GPA
- Highest level of math
- Grade in last math class
- Years since last math class
- Time of attendance
- Units planned
- Work hours planned
- Educational goal
- Definite major choice
- Importance of college to self
- Importance of college to others
- Parent's education

Many of these variables are based on research regarding student success and persistence (Nora & Rendon, 1990; Nettles et al., 1986). Though the use of these variables is seldom questioned, how to use them for prediction often is. In the California Community Colleges, assessment of students to help place them in their first semester courses is highly regulated. Part of that regulation is the requirement to use multiple measures, but how to use the measures and which measures to use is left to the discretion of individual colleges (California Community Colleges, 1998). In addition, if a test is used, it must meet strict requirements regarding validation; the overall placement process, too, must meet validation requirements. However, no requirement regarding the validation of multiple measures exists. This leads to the highly subjective use of multiple measures for placement as well as the common practice of collecting more data than is used for analysis.

Faced with this same dilemma, the initial purpose of this research was to utilize these data for placement. The intention was to build a model so that placement could be predicted using all these variables. To that end, I started to build a model using multiple regression. The initial model used these variables to predict success in three levels of English courses: college level (English 1A) and two levels below college level (English 1B and English 1C, respectively).

The use of multiple regression to build a model to predict future behavior has been utilized in education for a multitude of studies (Schroeder et al., 1986). The use of multiple regression for building a model is obvious: the computer output[1] includes both the standardized and unstandardized coefficients. The standardized coefficients give the relative importance of each variable while the unstandardized coefficients allows the creation of a model based on the coefficients. In addition, multiple regression handles the use of dichotomous dependent

[1] SPSS for Windows (Version 10.0.5) was used in these analyses.

variables effectively (Cohen & Cohen, 1983), as is the case in this research, specifically, success (A, B, C, or CR) or nonsuccess (D, F, I, or W).

Because of the large number of variables and the fact that there was no unified theory dictating the use of particular variables (Schroeder et al., 1986; Tabachnick & Fidell, 1989), the stepwise method of multiple regression was used for analysis. The resulting models for each of the courses are presented below in Table 1.

Insert Table 1 about here

Utilizing the unstandardized coefficients, I was able to build a model for predicting success in these three levels of English.

In preparing the report, however, I came upon a problem with this method of model building. Though the rationale and technique were acceptable, these models could not be used for placement. Why? Because the models were built to predict success in each course, not to predict which the course each student belonged. An example might help explain this shortcoming. If these models were going to be used to place students in an English course, which set of variables would be used? If the English course in which the student was to enroll was know, the various variables for that model could be employed to predict success in that course. Without knowing into which course the student was to enroll, these models were useless.

Discriminant Function Analysis

This led me to investigate the use of Discriminant Function Analysis to answer this question. Discriminant function analysis is a statistical technique used for classifying observations (Klecka, 1980). Some examples of research using this technique include predicting success in academic programs, identifying variables that to determine voting behavior, determining authorship of papers, or determining outcomes of terrorist hostage situations –

discriminant function analysis can be used in all of these examples (Klecka, 1980; Mosteller & Wallace, 1964).

As with any statistical technique, the proper use of the test requires that assumptions underlying the technique be observed (Klecka, 1980; Tabachnick & Fidell, 1989). The independent variables need to be interval while the dependent variable, the groups into which observations are classified, need to be nominal. Multivariate normality is assumed, but discriminant function analysis is robust to violations due to skewness rather than outliers (Tabachnick & Fidell, 1989). Discriminant function analysis does, however, include a technique that can be used to identify outliers, Mahalanobis distances, as a built-in option.[2] Homogeneity of variance-covariance matrices is another assumption of discriminant function analysis, but like multivariate normality, discriminant function analysis is robust to violations. Finally, violations of multicollinearity may make the underlying matrix calculations unstable and must be avoided but can be controlled with an option in the program. Generally, violations of these assumptions are conservative; that is, the power of the test is reduced, thereby lessening the chance of finding significance (Klecka, 1980).

Discriminant function analysis produces functions that help define the groups; the maximum number of functions that can be defined is one less than the number of groups. The functions first seek to distinguish the first group from the others, then the second group from the rest, and so on. These are identified by the Eigenvalues on the output. The eigenvalues also show what percent of variance is accounted for with each function. In addition, Wilks lambda tests the significance of each function.

---

[2] The technique for assessing and handling violations of assumptions is beyond the scope of this paper. The reader is directed to consult any of the several current books that deal with using statistical technique with various computer programs such as Tabachnick and Fidell (1995) or Klecka (1980).

For this research, the groups used in this analysis were defined as those who were successful in English 1A, English 1B, and English 1C. When discriminant function analysis was applied to these data to distinguish between these groups, it first identified a function that distinguished English 1A from the other two courses. Next, it identified a function to distinguish between English 1B and English 1C. The eigenvalues in Table 2 show that function 1 accounts for 95.4 percent of the variance while function 2 accounts for only 4.6 percent. The significance of Wilks lambda shows that both functions are statistically significant, so both can help distinguish between groups. However, it is easier to distinguish between English 1A and the other two courses than it is to distinguish between English 1B and English 1C.

Insert Table 2 about here

One of the benefits of discriminant function analysis is that it produces a classification table, showing where the data were categorized and in which groups they were predicted to be (see Table 3). The table includes the percent of cases correctly classified through the prediction of group membership. Since discriminant function analysis will classify cases into the largest group, a statistic, tau, can be computed showing the proportional reduction of error (PRE) when using the predicted model.

Insert Table 3 about here

To compute tau, subtract the percent of the largest group from the percent "correctly classified" as identified at the bottom of the classification table (see Table 3). Then divide this number by the percent of the largest group subtracted from 1. In this example, the percent correctly classified is 62.6% and the percent of the largest group is 55%. The PRE for this

research shows that placements based on this model increase by almost 17%, which translates into about 178 students placed more correctly using this method.

Discriminant function analysis output includes both standardized and unstandardized weights. The standardized weights show the relative importance of each variable compared to each other while the unstandardized weights show the relative significance of each variable based on its own scale of measurement. Table 4 below shows the standardized weights for the model. The variable, "Grade in last English class," has the greatest effect for predicting membership into group 1 than another other variable, followed by "Highest math class," though it has an inverse relationship to group membership. For distinguishing group 2 from group 3, the variable "Have a learning disability," is the single strongest predictor for membership in group 2 while the other variables have less significance.

Insert Table 4 about here

The structure matrix (Table 5) shows the how all the variables relate to each function at the same time. The output of discriminant function analysis illustrates that all the variables in the model predict group membership to some extent, even though small. Also, each variable contributes some amount to each group at the same time. However, the absolute value of its contribution helps determine to which group each variable belongs.[3] The SPSS output organizes the variables by group, listing the variables that contribute the most to group 1 first, then group 2.

Insert Table 5 about here

Despite all the output, I was once again faced with the problem of developing a predictive model based on discriminant function analysis. Upon further investigation, I found that there

[3] The superscripts of "a" denote that those variable were excluded from the final model based on stepwise discriminant function analysis.

BEST COPY AVAILABLE

were two basic methods of developing prediction. I could either compute variables using matrix algebra, or I could use Fisher's Linear Discriminant Functions.

The use of either method is basically the same. For each group, a function is computed for each case. For these data, three different functions result for each case. Whichever function is largest determines into which group that case is predicted to belong. I decided to use Fisher's Linear Discriminant Functions since the coefficients could be easily produced in the output and because the computation of linear function was easier than using matrix calculations. For each case, the response for each variable in the final model is multiplied by the coefficient produced by Fisher's Linear Discriminant Functions. Then, the products are added, resulting in linear composite for each case. For example, suppose that a student responded to the following questions with the following responses. Looking at Table 6, for group 1, sex would be multiplied by 7.9. Next, "ESL" (English as a Second Language) would be multiplied by 9.906 and so on. Next, each response would be multiplied by the coefficients in the second column and summed and then for the third column. The equations would be, respectively: 165.602, 165.665, 165.25. Since the highest sum is 165.665, the case would be predicted to be in group 2.

Insert Table 6 about here

As a check of these figures, I compared the predicted group membership using Fisher's Linear Discriminant Functions with that produced by the SPSS output and found that using this procedure produced the same group membership predictions as determined by SPSS.

Summary

The use of discriminant function analysis to classify data can be an extremely useful tool for researchers and college administrators. A plethora of data can be utilized simultaneously to classify cases and the resultant model can be evaluated for usefulness relatively easily. The

ability to develop a predictive model based on the model produced through the discriminant function analysis procedure increases its usefulness substantially. Colleges can utilize this dynamic and powerful procedure to target services and interventions to students who need it most, thereby utilizing their resources more effectively.

References

Anastasi, A. (1982). Psychological Testing. New York: Macmillan.

California Community Colleges (1998). Matriculation Regulations. Sacramento: California Community College Chancellor's Office, Student Services and Special Programs Division.

Cleary, T. A., Humphreys, C., Kendirick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. American Psychologist, 30, 15-41.

Cohen, J., & Cohen, P. (1983). Applied multiple regression /correlation analysis for the behavioral sciences. (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Ebmeir, H., & Schmulbach, S. (1989). An Examination of the selection practices used in the talent search program. Gifted Child Quarterly, 33, 134-141.

Ewell, P.T. In Diane F. Halpern (Ed.), Student outcomes assessment: what institutions stand to gain (pp. 9-24). New Directions in Higher Education, No. 59. Vol. XV, number 3. San Francisco: Josey-Bass.

Greene, J., & McClintock, C. (1985). Triangulation in evaluation: design and analysis issues. Evaluation Review, 9, 523-545.

Klecka, W. R. (1980). Discriminant analysis. Beverly Hills: Sage.

Markert, L.F., & Monke, R.H. (1990). Changes in counselor education admissions criteria. Counselor Education and Supervision, 30, 48-57.

Mathison, S. (1988). Why triangulate? Educational Researcher, 17, 13-17.

McNabb, S.L. (1990). The uses of "inaccurate" data: a methodological critique and applications of Alaska Native data. American Anthropologist, 92, 116-129.

Melnick, M. (1975). Testing the disadvantaged. American Psychologist, 30, 944-945.

Mosteller, F., & Wallace, D. L. (1964). Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley.

Nettles, M. T., Thoeny, A. R., & Gosman, E. J. (1986). Comparative and predictive analyses of black and white students' college achievement and experiences. Journal of Higher Education, 57(3), 289-318.

Nora, A., & Rendon, L. I. (1990). Determinants of predisposition to transfer among community college students. Research in Higher Education, 31(3), 235-255.

Riehl, R. J. (1994). The academic preparation, aspirations, and first-year performance of first-generation students. College and University, 70(1), 14-19.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). Understanding regression analysis: an introductory guide. Newbury Park: Sage.

Selingo, J., & Brainard, J. (2001, March 2). Call to eliminate SAT requirement may reshape debate on affirmative action. The Chronicle of Higher Education, pp. A21.

Stage, F.K., & Russell, R.V. (1992). Using method triangulation in college student research. Journal of College Student Development, 33, 485-491.

Sturm, S., & Guinier, L. (2001). The future of affirmative action. Boston Review, December 2000/January 2001, 4-10.

Tabachnick, B. G., & Fidell, L. S. (1989). Using multivariate statistics. (2nd ed.). New York: HarperCollins.

Tracey, T. J., & Sedlacek, W. E. (1985). The relationship of noncognitive variables to academic success: A longitudinal comparison by race. Journal of College Student Personnel, 26, 405-410.

Young, B.D., & Sowa, C.J. (1992). Predictors of academic success for Black student athletes. Journal of College Student Development, 33, 318-324.

Table 1: Predictive models for English1A, 1B, and 1C

Course	Variables in the final equation	Statistics
English 1A	High school GPA Age Sex Grade in last math class Ethnicity	R=.267, p<.05
English 1B	Highest level of math Grade in last English class Definite major choice Work hours planned	R=.273, p<.05
English 1C	Highest level of math	R=.603, p<.05

Table 2: Eigenvalues for discriminant functions

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.369	95.4	95.4	.519
2	.018	4.6	100.0	.132

a First 2 canonical discriminant functions were used in the analysis.

Table3: Classification Results

			Predicted Group Membership			Total
		Success in English 1A, 1B, or 1C	1.00	2.00	3.00	
Original	Count	English 1A	182	155	4	341
		English 1B	119	450	18	587
		English 1C	3	94	26	123
		Ungrouped cases	123	435	52	610
	%	English 1A	53.4	45.5	1.2	100.0
		English 1B	20.3	76.7	3.1	100.0
		English 1C	2.4	76.4	21.1	100.0
		Ungrouped cases	20.2	71.3	8.5	100.0

a 62.6% of original grouped cases correctly classified.

Table 4: Standardized Canonical Discriminant Function Coefficients

	Function 1	Function 2
Sex	.209	.284
English primary language	.305	.167
Have learning disability	-.204	.831
Admission status	-.269	-.133
Grade in last English class	.686	-.274
Highest math class	-.432	-.242
Grade in last math class	.227	.287
Educational goal	-.267	-.037

Table 5: Structure matrix

	Function 1	Function 2
Grade in last English class	.675	-.211
Highest math class	-.542	-.192
HS GPA[a]	.419	.058
Grade in last math class	.333	.270
Educational goal	-.266	-.046
Years since last math class[a]	.202	.014
English primary language	.200	.154
Years out of school[a]	.183	.070
Units planned[a]	-.153	-.034
Agea	.151	.041
Years of HS English[a]	-.129	-.039
Importance of college to others[a]	-.107	-.042
Definite major choice[a]	.063	-.046
Importance of college to self[a]	-.054	.038
Plan to attend[a]	.044	-.008
Highest college degree[a]	.033	.022
Veteran[a]	.032	.031
HS education[a]	.031	-.008
Have learning disability	-.168	.813
Sex	.133	.344
Admission status	-.126	-.132
Mothers education[a]	.038	.080
Work hours planned[a]	-.008	-.063
Fathers education[a]	-.004	-.051
Ethnicity[a]	-.009	.032

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

* Largest absolute correlation between each variable and any discriminant function

[a] This variable not used in the analysis.

Table 6: Fisher's linear discriminant functions (Classification Function Coefficients)

	Success in English 1A, 1B or 1C		
	English 1A	English 1B	English 1C
Sex	7.900	8.427	8.658
English primary language	9.906	10.963	11.817
Have learning disability	124.640	124.507	120.248
Admission status	1.445	1.220	1.031
Grade in last English class	1.851	2.602	3.676
Highest math class	1.556	1.256	1.015
Grade in last math class	2.155	2.450	2.591
Educational goal	10.471	10.092	9.695
(Constant)	-171.858	-171.666	-166.197

Fisher's linear discriminant functions